Advances in Computer Science and Information Technology (ACSIT)
Print ISSN: 2393-9907; Online ISSN: 2393-9915; Volume 2, Number 4; April-June, 2015 pp. 393-397

http://www.krishisanskriti.org/acsit.html

Data Driven Analytics: Prospects and Challenges

Suman Madan[1] and Aakanksha Chopra[2]

[1,2](IT),Jagan Institute of Management Studies (JIMS), Sec-5, Rohini, New Delhi- 110085
Affiliated to GGSIPU, Dwarka Sec-16 New Delhi-110078
E-mail: [1]madan.suman@gmail.com, [2]aakankshachopra.spm@gmail.com

Abstract—*Information generated by people and the gadgets today has gone from deficient to abundant which has increased the complexity of data. The challenge is how to elicitate, manage and organize this big data. The actual thing is to find out which data is important, what to keep and what to discard, and where to keep this enormous amount of data. The volume of data gets expanded when data is linked with other, resulting in data integration. Big data is still viewed as conventional data which is not sufficient. It should focus on what value the data can create into analytics rather than what technology it brings. Today due to digital convergence we have an opportunity and a challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. This paper focuses on big data analytics, the challenges and opportunities it accelerates and how it could be magnified to base the analytics.*

Keywords: *big data pipelining, big data analytics.*

1. INTRODUCTION

The potential of data-driven decision-making is accepted worldwide and there is growing buzz for the new discipline 'Big Data'. The age of big data has already begun. A joint study of 1144 IT and business professionals by IBM and the University of Oxford's said Business School found a 70% increase between 2010 and 2012 in those who reported that using big data afforded them a "competitive economic advantage". In 2014, worldwide data generation is estimated at a staggering 7ZB [1], and by 2018 only smart phones is expected to generate 2GB of data every month [2] since these smart phones are generating location and other data that keeps services running and ready to use. At the same time, the big data technology and services market is expected to grow at a 40 percent compound annual growth rate (CAGR) – about seven times the rate of the overall ICT market – with revenues expected to reach USD 16.9 billion in 2015 [3]. Also, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009, thus there is currently a huge gap between its potential and its realization. Scientific research has been revolutionized by Big Data [10].

Today our daily lifestyle creates huge digital record, may be sharing our thoughts and opinions on Facebook, twitter or other social media, streaming a video, playing the latest game with friends or sharing our photos. The decision makers of all industries, researchers and scientists, would like to base their decisions and actions on this data. The decisions may vary from designing more competitive offers, prices and packages; recommending the most attractive offers to subscribers during the shopping and ordering process; communicating with users about their usage, spending and purchase options; configuring the network to deliver more reliable services; and monitoring QoE to proactively correct any potential problems. All these activities enable improved user experience, increased loyalty, the creation of smarter networks, and extended network functionality to facilitate progress toward the Networked Society.

In 2010, the users and organizations stored around 13 Exabyte of new data which is over 50,000 times the data in the Library of Congress. The impending value of global personal location data for end-users is estimated to be $700 billion, and it can result in an up to 50% decrease in product development and assembly costs, according to a recent McKinsey report [11]. McKinsey predicts an equally great effect of Big Data in employment, where 140,000-190,000 workers with "deep analytical" experience will be needed in the US; furthermore, 1.5 million managers will need to become data-literate. Not surprisingly, the recent PCAST report on Networking and IT R&D[12]identified Big Data as a "research frontier" that can "accelerate progress across a broad range of priorities." Even popular news media now appreciates the value of Big Data as evidenced by coverage in the Economist [13], the New York Times [14] and National Public Radio [15, 16].

Big data analytics has the capacity to process any variety, volume and velocity of information and to derive an insight into data [5]. The 4 V's of big data: volume, variety, velocity, and veracity [6]:

1.1 Volume

This refers to amount of data. Since digital data is growing fast and data storage technologies improved, we can store larger amounts of data more cheaply. This resulted in analysis of even older data that was usually discarded.

1.2 Variety

This refers to forms of data. In the recent past, data largely existed in static spreadsheets, but now-a-days the data formats

that are more reflective of everyday activities. The data has shifted from structured like spreadsheets to unstructured like maps, video, email etc.

1.3 Velocity

This deals with speed of data i.e. "time between when data is created or captured, and when it is accessible". With improvements in technology, data can now be collected, shared, and analyzed more quickly and often in real time. This reduces the gap between when data is generated and when it becomes actionable in decision-making.

1.4 Veracity

This refers to reliability of data. Data quality is a challenge regardless of scale. However the increase in volume, variety, and velocity intensify this issue, promising datasets that need to be rigorously evaluated for their accuracy, origin, relevance, and consistency.

(See Fig. 1)Underlying the V's are significant shifts in technology development coupled with movements in business strategy from focusing on causal relationships to predictive analysis [7, 8, 9]. Earlier when data storage was expensive and size of storage limited, data was collected more selectively and decisions were made prior to the collection process. With technological advancements, the storage capacities expanded and simultaneously became less expensive, larger datasets could be collected and stored, allowing for more options and flexibility in analysis. Consequently, datasets continue to expand. As Eric Schmidt, Executive Chairman of Google famously said in 2010, "There was 5 Exabyte of information created between the dawns of civilization through 2003, but that much information is now created every 2 days, and the pace is increasing [r4]."

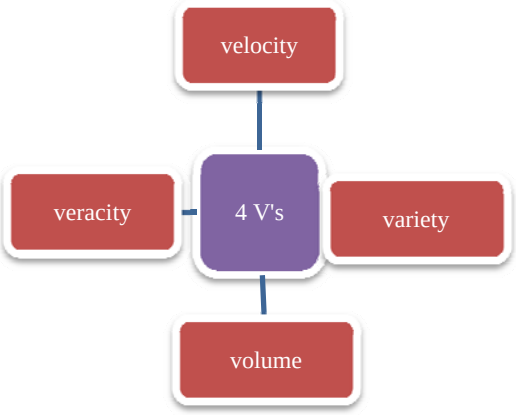

Fig. 1: (a): 4 V's of Big Data

Getting value out of big data is no longer a illusion. However for decision makers doing network near analytics, extraction of valuable data is a domain-specific task, and off-the-shelf IT components are not good enough. Bringing together large datasets allows for matching and connections that were not previously possible. Therefore, decision makers need to look beyond traditional big data techniques and focus on the analytic value that can be gained from transforming the 4 V's into 3 A's i.e. into pure business insight: Adequate, Accurate and Actionable. These A's enables the decision makers to improve their existing process and drive better decision-making. These insights provides following benefits to organizations:

i. Operation innovation leading to business innovation
ii. Operation efficiency leading to business efficiency
iii. User satisfaction leading to Customer personalization and retention.

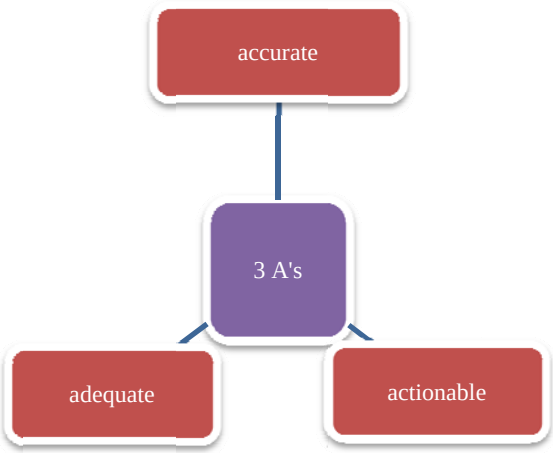

Fig. 1: (b): 3 A's of Big Data

2. UNDERSTANDING THE BUSINESS OF BIG DATA

As such there is no blueprint which can convert the usage of big data into money. With innovative uses constantly emerging and few success stories of big organizations, following two major dimensions can be considered that organizes practices into key decision stages at which critical choices about data use affect costs and profit:

a. Data types and sources
b. Big data business models

2.1 Data types and sources

Data are defined by how they are acquired as it can be drawn from multiple sources. While sectors such as retail may rely on transactional data more heavily than the health or insurance sectors, they share a dependence on diverse datasets from a mix of open and proprietary sources.

Often, data is combination of transactional data collected in-store and online, demographic data, and geographic data, among other data sources. This data can be collected from in-store campaigns, promotions sent via email or text to customers, professional, and social media sites, satellite imagery, river gauging stations, earthquake sensors, and storm forecasts to analyze risks presented by catastrophic events.

Therefore, the question arises: *"How can one categorize these varied sources of data? "*. The diverse and evidently complex, experts from various fields like finance, retail, media, and advertising describe data in three categories according to acquisition points and ownership:

2.1.1 First party

It means data owned by the business, collected about a business' customers, for example, transactional data, demographic data collected directly from customers generally for commercial purposes, for either in-house analysis or sale to third parties. This data is closed. It is not publicly shared or freely available. First party data can be viewed as an output that may have value in its own right and can serve as a basis for proprietary competitive advantage.

2.1.2 Second party

It means derived data where ownership is uncertain like data collected in collaboration with another company. This data are shared between companies as per contractual agreement specifying a scope of time and behaviors. Ownership of this data is uncertain but generally the company collecting the data stores it. For example, Google AdWords data, how many page views, how many clicks, how many times a particular search term was entered. The nature of this data is also typically closed.

2.1.3 Third party

It means data collected and owned by different parties. For example: government datasets, credit scores and audience ratings. This data includes open or public data as well as data of a closed or commercial nature collected from the web thru web scraping or data collected and processed by data vendors. Third party data vendors package data according to client demand, providing information on specific target groups. Third party data is an often costly input. Business models then revolve around how to use such an input to create new products or enhance business efficiency in novel and innovative ways.

2.2 Big data business models

The business of big data is complex, with most companies engaging in multiple dimensions of collection, processing, and sale of data. Emerging in this market are "as a service" models in which companies provide software, platforms, analytics, or consulting "as a service," engaging their customers in multiple ways.

2.2.1 Monetizing first party data

In the most advanced organizations, first party data is used to inform internal business decisions on an extremely fine-grained scale. These data also inform decisions about products, pricing, promotions, stock keeping, and overall business strategy. While the primary business model for these companies is retail, first party data informs almost every significant decision. One obvious way to monetize proprietary first party data is to treat it like any other product and sell it to other parties. Thus, first party data is treated as an output in its own right.

2.2.2 Data analytics as a service

The value of data lies in the actions resulting from analysis and not in its intrinsic merits. Thus, a common business model is created for companies in the big data sphere are the provision of ***analytics as service*** where data analytics is part of the business model. In this model the analytics firm takes as an input its own proprietary data, data supplied by its client, or some third party source of data, and produces as an output a data summary, analysis, insight, advice, or some other product derived from that data. Analytics often result in reporting insights on a client's targeted audience *segment* based on aggregated behavioral data for particular groups. Since processing and analysis technologies are becoming less expensive and the consumers are more frequently using huge variety of data, big data analytics are also rapidly becoming available for personal use. Thus, a new variety of consumer-facing analytics firms is emerging. For example: Mappings, a mobile application that allows users to report their levels of happiness and receive feedback, collects personal data, analyses it, and reports it in a usable form to users.

2.2.3 Three distinct classes of big data business models-

a. Data users
These are organizations that use data either to inform business decisions, or as an input into other products and services. Such organizations faces questions like what data to be created, what data is needed externally, and how can this data be used to create value within the business? These businesses require the physical and human resources to take advantage of the data.

b. Data suppliers
These are organizations that either generate data that is of native value and therefore marketable or else serve a kind of brokerage role by providing access to an aggregation of first and third party data. Such organizations need not specialize in the supply of data and many organizations are finding that they hold data that is of considerable value when some third party puts it to a use other than that for which is was originally collected. The key questions are what data is available, what uses might that data have and for whom, and how should data be delivered to maximize its value?

c. Data facilitators
These organizations perform a range of services including advice on how to take advantage of big data, the provision of physical infrastructure, and the provision of outsourced analytics services. These organizations are playing important role during the current time of transition when a large number of firms are reorganizing to make data more central to their

business, but still lack the internal expertise to do so without assistance.

3. BARRIERS OF USING BIG DATA

Big data promises great potential but using big data is a cluttered process involving extensive cleaning and normalizing of data. Furthermore, there is the problem of missing data and the challenge of enumerate phenomenon such as risk may defy measurement. Some common challenges that impede progress and underlie many and sometimes all phases of big data pipeline: *Heterogeneity, scale, timeliness, privacy, human collaboration and complexity problems* (see Fig. 2). Following are the challenges in every phase of big data pipeline that potentially affect productivity and profitability or may create barriers for businesses, expected or otherwise.

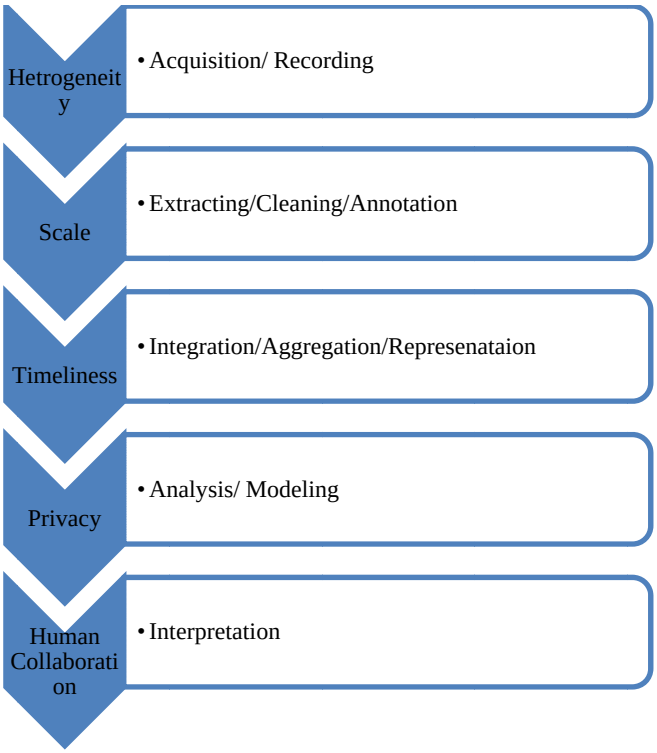

Fig. 2: Pipeline showing barriers in using Big Data

3.1 Data Acquisition and Recording

The first phase challenge is finding the appropriate data source since data is available in wide variety of sources and formats. Many times the data is available online but decision makers had to struggle out on what data is available and in which format. Generally machine readable data is available in schema format or in PDF files filled with text. Thus data analyst has to filter and compress available huge datasets by orders of magnitude. **Challenges** can be:

a. Define these filters in such a way that they do not discard useful information.
b. Data that is spatially and temporally correlated needs to be intelligently processed and reduced so that loss doesn't happen
c. Finding "on-line" analysis techniques which can process continuously streaming data as it is difficult to store first and reduce afterward.
d. Automatically generating the right metadata to describe what data is recorded and how it is recorded and measured.
e. Data provenance is another important issue. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. With suitable provenance, we can easily identify all subsequent processing that dependent on this step. Thus we need research both into generating suitable metadata and into data systems that carry the provenance of data and its metadata through data analysis pipelines.

3.2Information Extraction and Cleaning

The information collected will not be in a format ready for analysis. Thus, an information extraction process is required to extract the required information from the underlying sources and expresses it in a structured form suitable for analysis. **Challenges** are:

a. Extracting correct and complete information is a continuous challenge.
b. Data cleaning requires well-recognized constraints on valid data or well-understood error models.
c. Since cleaning is a process that occurred only after code, the analysts needs to capture the justifications for cutting irregular data points: How had they detected the error? What model had found the issue?

3.3 Data Integration, Aggregation, and Representation

Heterogeneity, scale, timeliness, complexity, and privacy problems with Big Data impede progress at all phases of the pipeline that can create value from data. **Challenges** are:

a. Since there are differences in data structure and semantics of collected data, it needs to be expressed in forms that are computer understandable and resolvable.
b. Additional work in data integration is required to achieve automated error-free difference resolution.
c. What is the suitable database design? For effective database designs, creating tools and techniques for design process so that databases can be used effectively in the absence of intelligent database design.

3.4 Query Processing, Data Modeling, and Analysis

Big data is dynamic, diverse, deceitful and noisy in nature but could be more valuable because statistics obtained from frequent patterns and correlation analysis usually overcome individual fluctuations and often unveil more reliable hidden patterns and knowledge. Data Mining requires integrated, cleaned, trustworthy, and efficiently accessible data,

declarative query and mining interfaces, scalable mining algorithms, and big-data computing environments. Challenges are:

a. Creating automated Queries towards Big Data for content creation on websites, to populate hot-lists or recommendations, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it.
b. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major challenge.
c. Establishing the missing coordination between database systems that host the data along with SQL querying procedures and with analytics packages that perform various forms of non-SQL processing, such as data mining and statistical analyses. This will benefit both expressiveness and performance of the analysis.

3.5 Interpretation

Once the effective analysis is done, the decision maker must interpret the analysis results which needs investigation of all the assumptions made and retracing the analysis. Challenges are:

a. Errors can come in many forms: bugs in systems, models based on assumptions, results based on erroneous data. Thus decision makers don't totally trust computer system and try to verify results themselves. This poses major challenge with Big Data due to its complexity.
b. Users need to be able to see not just the results, but also understand how best to capture, store, and query provenance along with techniques to capture adequate metadata. This is too technical and many users to don't grasp this completely.

4. CONCLUSION

This era belongs to big data and big data analytics is an emerging type of knowledge work offering plenty of opportunities for study and productivity improvements. However, many technical challenges discussed in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. However, these challenges will require transformative solutions and following **recommendations** are suggested: Data must be central to business model, profit model should be clear, Strong business strategy paired with understanding of technology and Look for Low-Hanging Fruit.

REFERENCES

[1] Villars, R. L., Olofson, C., W., and Eastwood, M., "Big Data: What it is and Why You Should Care", *IDC Analyze the Future,* June 2011, Available at: http://sites.amd.com/us/Documents/IDC_AMD_Big_Data_Whitepaper.pdf

[2] Ericson Mobility Report- On the pulse of Networked Society, *Ericson 2013*, June 2013 Available at: http://www.ericsson.com/res/docs/2013/ericsson-mobility-report-june-2013.pdf

[3] **"**Worldwide Big Data Technology and Services 2012-2016 Forecast", *IDC Press Release*, 2012 Available at: http://laser.inf.ethz.ch/2013/material/breitman/additional%20reading/Worldwide%20Big%20Data%20Technology%20and%20Services%202012-2016%20Forecast.pdf

[4] http://techonomy.typepad.com/blog/2010/08/google-privacy-and-the-new-explosion-of-data.html

[5] Corrigan, D., "Big Data: Achieving Competitive Advantage through Analytics", *Big Data Integration and Governance, IBM*,2012 IBM, Available at: https://www-950.ibm.com/events/wwe/grp/grp037.nsf/vLookupPDFs/Calgary_Keynote_%20David_%20Corrigan%20-%20v1/$file/Calgary_Keynote_%20David_%20Corrigan%20-%20v1.pdf

[6] Schroeck, M., Shockley, R., Smart, J., Romero-Morales, D., and Tufano, P., "Analytics: The Real-World Use of Big Data",London: IBM Global Business Services Business Analytics and Optimization in collaboration withSäid Business School, University of Oxford, 2012 IBM, Available at: http://www-03.ibm.com/systems/hu/resources/the_real_word_use_of_big_data.pdf

[7] Gild. (2013). "The Big Data Recruiting Playbook". San Francisco: Gild.Retrieved from http://www.gild.com/resource/big-data-recruiting-playbook-2/

[8] O'Neil, C., (2014),"On Being a Data Skeptic", *Printed in the United States of America, Published by: O'Reilly Media, Inc., California, 1005 Gravenstein Highway North, Sebastopol, CA 95472*, 2014, Available at: http://www.oreilly.com/data/free/files/being-a-data-skeptic.pdf

[9] Mayer-Schönberger, V. & Cukier, K., (2013), "Big Data: A Revolution that will Transform How We Live, Work, and Think*"New York: Houghton Mifflin Harcourt Publishing Company.*

[10] Advancing Discovery in Science and Engineering. Computing Community Consortium. Spring 2011.[CCC2011a]

[11] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., and Byers, A., H., "Big data: The next frontier for innovation, competition, and productivity", *McKinsey Global Institute. May 2011.*[McK2011]

[12] "Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology." *PCAST Report*, Dec. 2010. Available at: http://www.whitehouse.gov/sites/default/files/microsites/ostp/pcast-nitrd-report-2010.pdf

[13] "Drowning in numbers -- Digital data will flood the planet—and help us understand it better", *The Economist*, Nov 18, 2011.Available at: http://www.economist.com/blogs/dailychart/2011/11/big-data-0

[14] Lohr, S.,"The Age of Big Data", *New York Times*, Feb 11,2012. Available at: http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html

[15] Noguchi, Y., "Following the Breadcrumbs to Big Data Gold", *National Public Radio*, Nov. 29, 2011. Available at: http://www.npr.org/2011/11/29/142521910/the-digital-breadcrumbs-that-lead-to-big-data

[16] Noguchi, Y., "The Search for Analysts to Make Sense of Big Data", *National Public Radio*, Nov. 30, 2011.Available at: http://www.npr.org/2011/11/30/142893065/the-search-for-analysts-to-make-sense-of-big-data